Economic Clustering Summary Report: South Africa 2016

Christelle Nyoumbi

01 August, 2023

# Summary of Data

**Country Code-year:** ZA16

**Number of observations:** 11083

**Number of clusters:** 5

**Number of variables used:** 23

**Distance used:** Hamming

**Variables used in the algorithm:** hv201a, hv206, hv207, hv208, hv209, hv212, hv225, hv237, hv243b, hv243e, hv246, hv252, sh121g, sh121h, sh121i, sh121j, sh124b, water, toilet, floor, roof, cookfuel, wall

Using the hamming distance as a metric and the silhouette score as a measure of the homogeneity of a partition, a classification is performed, the results of which are analyzed here. The value of a partition’s silhouette score lies between 0 and 1. The greater its value, the better the grouping of points in the partition.

# Top Variable Groupings (Sorted by ASW)

| Group | ASW | Var. 1 | Var. 2 | Var. 3 | Var. 4 |
| --- | --- | --- | --- | --- | --- |
| 1 | 0.9411 | has electricity | has refrigerator | vacuum cleaner or floor polisher | washing machine |
| 2 | 0.9408 | has electricity | has refrigerator | vacuum cleaner or floor polisher | microwave oven |
| 3 | 0.9376 | has electricity | has a computer | vacuum cleaner or floor polisher | roof |
| 4 | 0.9375 | has electricity | vacuum cleaner or floor polisher | microwave oven | electric/gas stove |
| 5 | 0.9369 | has electricity | vacuum cleaner or floor polisher | electric/gas stove | roof |
| 6 | 0.9364 | has electricity | has a computer | washing machine | roof |
| 7 | 0.9358 | has refrigerator | vacuum cleaner or floor polisher | electric/gas stove | roof |
| 8 | 0.9355 | has electricity | has a computer | microwave oven | roof |
| 9 | 0.9355 | has electricity | microwave oven | washing machine | roof |
| 10 | 0.935 | has electricity | has refrigerator | vacuum cleaner or floor polisher | roof |

This table presents a classification into 10 classes, indicating the silhouette score for each class and the 04 most important variables (in order of importance: var 1, var 2, var 3, var 4). ASW, which is a measure of how similar individuals within a cluster are to each other and how dissimilar they are to individuals in neighboring clusters indicates strong clusters,i.e highly homogeneous partitions (silhouette score > 0.9 for all partitions).

Group 1 and Group 2 have the highest value of 0.94, suggesting a better distinction from the other groups.

# 

# Marginal Distributions

| Variable | Description | % Time in Top Clusters | Distribution |
| --- | --- | --- | --- |
| hv206 | has electricity | 90% | Binary, 90% 1s (or yes) |
| hv209 | has refrigerator | 40% | Binary, 74.5% 1s (or yes) |
| hv243e | has a computer | 30% | Binary, 19.6% 1s (or yes) |
| sh121g | vacuum cleaner or floor polisher | 70% | Binary, 12.2% 1s (or yes) |
| sh121h | microwave oven | 40% | Binary, 49% 1s (or yes) |
| sh121i | electric/gas stove | 30% | Binary, 83.2% 1s (or yes) |
| sh121j | washing machine | 30% | Binary, 34.8% 1s (or yes) |
| roof | Roof | 70% | Categorical, 0 (finished) = 95.6%,   1 (natural) = 2.2%,   2 (other) = 0.1%,   3 (rudimentary) = 2% |

This table provides a description of each variable (name, designation, cluster generation time for each variable - as a percentage of the maximum time, variable type and value distribution).

Variables are ordered according to the percentage of time they spend in the best groups obtained previously. This means that the most discriminating variables, those most representative of the best groups, appear first in the table. Variables with a higher percentage of time spent in the best groups are considered more important in distinguishing the groups. Consequently, the variables that appear at the top of the table are considered the most discriminating.

# 

# Cluster #1 Configuration

| Cluster Group | has electricity | has refrigerator | vacuum cleaner or floor polisher | washing machine | Proportion (%) |
| --- | --- | --- | --- | --- | --- |
| 1 | 1 | 0 | 0 | 0 | 15.37 |
| 1 | 1 | 0 | 1 | 0 |  |
| 2 | 0 | 0 | 0 | 1 | 23.85 |
| 2 | 0 | 1 | 0 | 1 |  |
| 2 | 1 | 0 | 0 | 1 |  |
| 2 | 1 | 1 | 0 | 1 |  |
| 3 | 0 | 0 | 0 | 0 | 9.16 |
| 4 | 0 | 1 | 1 | 1 | 10.91 |
| 4 | 1 | 0 | 1 | 1 |  |
| 4 | 1 | 1 | 1 | 1 |  |
| 5 | 0 | 1 | 0 | 0 | 40.71 |
| 5 | 1 | 1 | 0 | 0 |  |
| 5 | 1 | 1 | 1 | 0 |  |

The table above describes the cluster named cluster 1 obtained following a 5-class classification partition. Cluster 1 is described by 5 sub-classes, themselves identifiable by patterns according to the values of the chosen variables (has electricity, has refrigerator, vacuum cleaner, washing machine). The proportion column indicates the proportions of each pattern.

Each row of the table represents a specific combination of variables with its corresponding proportion in the cluster group. The clusters are presented in numerical order, and empty cells indicate that the specific combination of variables is not present in the corresponding cluster. Thus, cluster 1 is made up of 5 types of individuals: - (has electricity, vacuum cleaner): 15.37% of the cluster 1 population - (has electricity, has refrigerator, washing machine) : 23.85 % of cluster population - (has electricity, has refrigerator, vacuum cleaner, washing machine): 10.91% - (has electricity, has refrigerator and washing machine): 40.71 % - No electricity, no refrigerator, no vacuum cleaner, no washing machine : 9.16%

# 

# Validation Tables

## a.1) Using Children Deceased (Sorted by proportion of 0%)

| Cluster ID/Children Deceased | 0% | 1-33% | 34-66% | 67+% |
| --- | --- | --- | --- | --- |
| 1\* | 630 (95.9%) | 14 (2.1%) | 12 (1.8%) | 1 (0.2%) |
| 2 | 2,231 (93.6%) | 82 (3.4%) | 54 (2.3%) | 17 (0.7%) |
| 5 | 3,399 (91.6%) | 169 (4.6%) | 103 (2.8%) | 38 (1.0%) |
| 4 | 907 (90.1%) | 48 (4.8%) | 36 (3.6%) | 16 (1.6%) |
| 3 | 665 (87.8%) | 60 (7.9%) | 24 (3.2%) | 8 (1.1%) |
| Total | 7,832 (92.0%) | 373 (4.4%) | 229 (2.7%) | 80 (0.9%) |
| \*The chi-squared p-value is 0 | | | | |

After partitioning into 5 classes, this table shows, for each class, the proportion of living children (the second column, 0%) and the cumulative proportions (of child death rates) of 1-33% (3rd column), 34-66% (4th column), and 67-100% (5th column). We have 7.832 children’s living according to different clusters and 682 of children deceased. If p is the p-value, 1-p represents the significance level. The chi-squared test was used to assess the association between clusters and child deaths. The p value 0 indicates a significant association between clusters and child deaths. This suggests that clusters are different in terms of child deaths.

Cluster 1 has the highest proportion of living children among the other clusters, which may be characterized by relatively favorable socio-economic conditions, contributing to a higher child survival rate.

Cluster 4: This cluster has a lower proportion of non-fatal children than the previous clusters(cluster 5). The proportion of deceased children is relatively high. This suggests a significant deterioration in socio-economic conditions, which could have a negative impact on child survival.

Cluster 3: This cluster has the lowest proportion of non-fatal children of all the clusters. The majority of children in this cluster have died, with a very low proportion of children who have not died. This suggests a very precarious socio-economic situation, with adverse conditions that can have a significant impact on child survival.

According to the different clusters, poorer socio-economic conditions are associated with lower child survival rates..

## a.2) Aggregating proportions greater than 0%

| Cluster ID/Children Deceased | 0% | >0% |
| --- | --- | --- |
| 1\* | 630 (95.9%) | 27 (4.1%) |
| 2 | 2,231 (93.6%) | 153 (6.4%) |
| 5 | 3,399 (91.6%) | 310 (8.4%) |
| 4 | 907 (90.1%) | 100 (9.9%) |
| 3 | 665 (87.8%) | 92 (12.2%) |
| Total | 7,832 (92.0%) | 682 (8.0%) |
| \*The chi-squared p-value is 0 | | |

# 

## b) Using Individual Education Level Attained (Sorted by weighted average by row)

| Cluster ID/Education | 0 | 1 | 2 | 3 | 4 | 5 | W. Avg. |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1\*a | 2 (0.3%) | 5 (0.8%) | 6 (0.9%) | 192 (29.2%) | 235 (35.8%) | 217 (33.0%) | 3.98 |
| 2 | 34 (1.4%) | 83 (3.5%) | 53 (2.2%) | 1,145 (48.0%) | 724 (30.4%) | 345 (14.5%) | 3.46 |
| 5 | 69 (1.9%) | 227 (6.1%) | 145 (3.9%) | 2,042 (55.1%) | 937 (25.3%) | 289 (7.8%) | 3.19 |
| 4 | 35 (3.5%) | 108 (10.7%) | 68 (6.8%) | 586 (58.2%) | 189 (18.8%) | 21 (2.1%) | 2.84 |
| 3 | 50 (6.6%) | 101 (13.3%) | 66 (8.7%) | 444 (58.7%) | 87 (11.5%) | 9 (1.2%) | 2.59 |
| Total | 190 (2.2%) | 524 (6.2%) | 338 (4.0%) | 4,409 (51.8%) | 2,172 (25.5%) | 881 (10.3%) | 3.23 |
| \*The chi-squared p-value is 0 | | | | | | | |
| a0=none,1=incomplete primary, 2=primary, 3=incomplete secondary, 4=secondary, 5=higher | | | | | | | |

Cluster 1: This cluster has a high proportion of individuals with education levels 4 (secondary) and 5 (higher). The level of education of individuals in this cluster is generally higher than in other clusters, with a weighted average of 3.98. This suggests that this cluster is characterized by a population with a relatively high level of education.

Cluster 3: In this cluster, the distribution of education levels is more balanced, with a similar proportion of individuals with education levels ranging from 1 (incomplete primary) to 4 (secondary). This cluster has the highest proportion of individuals with an education level of 0 (none). The weighted average education level in this cluster is 2.59, indicating a relatively low level of education..

Overall, there is significant variation in education levels between clusters. Clusters 1 and 2 have higher levels of education, while clusters 4 and 5 have lower levels.

# 

## c) Using Primary Healthcare Source (Sorted by % enrolled in public healthcare [ascending order])

| Cluster ID/Primary Healthcare Source | 0 | 1 | 2 |
| --- | --- | --- | --- |
| 1\*a | 149 (50.9%) | 120 (41.0%) | 24 (8.2%) |
| 2 | 885 (79.9%) | 127 (11.5%) | 96 (8.7%) |
| 5 | 1,500 (85.6%) | 131 (7.5%) | 122 (7.0%) |
| 4 | 437 (89.5%) | 17 (3.5%) | 34 (7.0%) |
| 3 | 336 (94.4%) | 4 (1.1%) | 16 (4.5%) |
| Total | 3,307 (82.7%) | 399 (10.0%) | 292 (7.3%) |
| \*The chi-squared p-value is 0 | | | |
| a0=public/government, 1=private, 2=other | | | |

The table is structured with clusters (Cluster ID) as rows and primary healthcare source categories (Primary Healthcare Source) as columns.

Cluster 1: Category 0 (primary care source “public/government”) is the most frequent in this cluster, accounting for 50.9% of observations. Categories 1 (primary care source “private”) and 2 (primary care source “other”) are also present, with proportions of 41.0% and 8.2% respectively.

Cluster 2: The three categories of primary healthcare source are fairly evenly represented in this cluster. Category 0 has the highest proportion with 79.9%, closely followed by category 1 with 11.5%, and category 2 with 8.7%.

Cluster 5: In this cluster, category 0 is the most frequent with a proportion of 85.6%. Category 1 accounts for 7.5% of observations, while category 1 is less frequent at 7%.

Cluster 3: Category 0 is the most frequent in this cluster, accounting for 94.4% of observations. Categories 1 and 2 have relatively lower proportions, at 1.1% and 4.5% respectively.

## Analysis of Maternal mortality variable MM9

Indicates if the respondent’s sister was pregnant when she died, if she died during childbirth, within six weeks after the delivery or within 2 months after the delivery. We have 20 sub-variables of MM9, from mm9\_01 to mm9\_20.

## Calculating the percentage of sibling’s death and pregnancy for mm9\_01

|  |  |  |  |
| --- | --- | --- | --- |
| death not related | 338 | 90.616622 | 90.61662 |
| died while pregnant | 14 | 3.753351 | 94.36997 |
| died during delivery | 12 | 3.217158 | 97.58713 |
| 2 months after delivery | 9 | 2.412869 | 100.00000 |
| Total | 373 | 100.000000 | 200.00000 |

This table shows that 90.62% of sibling’s death are not related to pregnancy and 19.38% (100 - 90.62) of sibling’s death are related to maternal deaths.

## d) Maternal Mortality data [Ascending order])for mm9\_01

| Cluster ID/Maternal Mortality | Not\_related | MM\_Rate |
| --- | --- | --- |
| 1\*a | 1.0000000 (100.0%) | 0.00000000 (0.0%) |
| 3 | 0.9736842 (97.4%) | 0.02631579 (2.6%) |
| 5 | 0.9036145 (90.4%) | 0.09638554 (9.6%) |
| 2 | 0.8947368 (89.5%) | 0.10526316 (10.5%) |
| 4 | 0.8571429 (85.7%) | 0.14285714 (14.3%) |
| \*The chi-squared p-value is 0 | | |
| aNot\_related=Death not related to pregnancy, MM\_Rate=Maternal Mortality Related | | |

This ranking reveals that there is no pregnancy-related mortality in cluster 1 and in other cluster it is lowest.

## Calculating the percentage of sibling’s death and pregnancy for mm9\_02

| sibling\_death\_and\_pregnancy | Freq | perc | cum\_perc |
| --- | --- | --- | --- |
| death not related | 229 | 91.235060 | 91.23506 |
| died while pregnant | 6 | 2.390438 | 93.62550 |
| died during delivery | 8 | 3.187251 | 96.81275 |
| 2 months after delivery | 8 | 3.187251 | 100.00000 |
| Total | 251 | 100.000000 | 200.00000 |

This table shows that 91.23% of sibling’s death are not related to pregnancy and 8.77% (100 - 91.23) of sibling’s death are related to maternal deaths.

## e) Maternal Mortality data [Ascending order]) for mm9\_02

| Cluster ID/Maternal Mortality | Not\_related | MM\_Rate |
| --- | --- | --- |
| 3\*a | 0.9583333 (95.8%) | 0.04166667 (4.2%) |
| 1 | 0.9411765 (94.1%) | 0.05882353 (5.9%) |
| 2 | 0.9230769 (92.3%) | 0.07692308 (7.7%) |
| 5 | 0.8983051 (89.8%) | 0.10169492 (10.2%) |
| 4 | 0.8888889 (88.9%) | 0.11111111 (11.1%) |
| \*The chi-squared p-value is 0 | | |
| aNot\_related=Death not related to pregnancy, MM\_Rate=Maternal Mortality Related | | |

This ranking reveals that pregnancy-related mortality is lowest in cluster 3, followed by cluster 1, cluster 2, 5 and lastly cluster 4.

## Calculating the percentage of sibling’s death and pregnancy for mm9\_05

| sibling\_death\_and\_pregnancy | Freq | perc | cum\_perc |
| --- | --- | --- | --- |
| death not related | 58 | 92.063492 | 92.06349 |
| 2 months after delivery | 5 | 7.936508 | 100.00000 |
| Total | 63 | 100.000000 | 200.00000 |

## 

## f) Maternal Mortality data [Ascending order]) for mm9\_05

| Cluster ID/Maternal Mortality | Not\_related | MM\_Rate |
| --- | --- | --- |
| 1\*a | 1.0000000 (50.0%) | 1 (50.0%) |
| 2 | 1.0000000 (50.0%) | 1 (50.0%) |
| 3 | 1.0000000 (50.0%) | 1 (50.0%) |
| 4 | 1.0000000 (50.0%) | 1 (50.0%) |
| 5 | 0.8571429 (46.2%) | 1 (53.8%) |
| \*The chi-squared p-value is 0 | | |
| aNot\_related=Death not related to pregnancy, MM\_Rate=Maternal Mortality Related | | |

## Calculating the percentage of sibling’s death and pregnancy for mm9\_06

| sibling\_death\_and\_pregnancy | Freq | perc | cum\_perc |
| --- | --- | --- | --- |
| death not related | 26 | 86.666667 | 86.66667 |
| died while pregnant | 2 | 6.666667 | 93.33333 |
| died during delivery | 2 | 6.666667 | 100.00000 |
| Total | 30 | 100.000000 | 200.00000 |

## g) Maternal Mortality data [Ascending order]) for mm9\_06

| Cluster ID/Maternal Mortality | Not\_related | MM\_Rate |
| --- | --- | --- |
| 2\*a | 1.0000000 (100.0%) | 0.0000000 (0.0%) |
| 4 | 1.0000000 (100.0%) | 0.0000000 (0.0%) |
| 5 | 0.8461538 (84.6%) | 0.1538462 (15.4%) |
| 1 | 0.5000000 (50.0%) | 0.5000000 (50.0%) |
| 3 | 0.0000000 (0.0%) | 1.0000000 (100.0%) |
| \*The chi-squared p-value is 0 | | |
| aNot\_related=Death not related to pregnancy, MM\_Rate=Maternal Mortality Related | | |